A SEQUENTIAL FEATURE SELECTION ALGORITHM FOR GMM-BASED SPEECH QUALITY ESTIMATION

Tiago H. Falk and Wai-Yip Chan

Department of Electrical and Computer Engineering
Queen's University, Kingston, Ontario, Canada
Email: {falkt, chan}@ee.queensu.ca

ABSTRACT

We propose a sequential feature selection algorithm for designing Gaussian mixture model (GMM) based estimators. Feature selection is performed progressively to minimize estimation errors. The algorithm is applied to design estimators of subjective speech quality. Simulation shows that estimators designed using the proposed algorithm outperform two benchmark algorithms by as much as 39% in correlation and 24% in root-mean-squared error. Furthermore, features selected by the proposed algorithm are suitable for diagonal GMM estimators, which incur lower computational complexity.

1. INTRODUCTION

In regression, given the values of n predictor variables, the value of a target variable is estimated by means of a mapping from the predictor variables to the target variable. Regression analysis provides a means to find a best mapping in the form of a regression function. This paper focuses on the case where the joint distribution of the predictor and target variables is modelled by a Gaussian mixture model (GMM). A large class of probability densities can be approximated using Gaussian mixtures. Moreover, GMMs provide a closed form expression for the regression function.

Estimation (or regression) using GMMs is introduced in [1] and the idea is used in [2] to adjust the magnitude spectrum of a speech signal when the fundamental frequency of the signal is altered. In [3] GMMs are used to estimate missing line spectral frequencies and in [4], subjective speech quality ratings.

The choice of the predictor or feature variables is often crucial in regression analysis, as redundant or noisy features degrade estimation performance. The problem at hand is to pick m feature variables out of $n > m$ variables for the regression function. The best m is often not known *a priori*, and an exhaustive search for an optimal feature subset entails examining $2^n - 1$ possible subsets, a clearly impossible task for large n. One approach is to use common feature selection algorithms such as classification and regression trees (CART) [5] and multivariate adaptive regression splines (MARS) [6].

When designing GMM-based estimators, the features selected by the above algorithms may not lead to high estimation accuracy. In [7], the concept of *feature saliencies* in the context of GMMs is proposed. By adopting a penalty criterion, saliencies of irrelevant features go to zero, thus performing feature selection. This feature selection procedure, however, does not take the GMM regression function into consideration and may still lead to features that are inefficient for the estimation task at hand.

Here, we propose a feature mining algorithm targeted to estimation tasks that make use of GMM regression functions. An experiment consisting of predicting the subjective quality rating of speech is performed and simulation results show that GMM estimators designed using our proposed algorithm outperform estimators trained on features selected by CART or MARS.

2. GAUSSIAN MIXTURE MODELS

2.1 Background

Let $\mathbf{u}$ be an K-dimensional vector, a Gaussian mixture density is a weighted sum of M component densities

$$p(\mathbf{u}|\mu,\mathbf{\Sigma},\alpha) = \sum_{i=1}^{M} \alpha_i . b_i(\mathbf{u}) \qquad (1)$$

where $\alpha_i \geq 0, i = 1,...,M$ are the mixture weights, with $\sum_{i=1}^{M} \alpha_i = 1$, and $b_i(\mathbf{u})$, $i = 1,...,M$ are the K-variate Gaussian densities with mean vector μ_i and covariance matrix $\mathbf{\Sigma}_i$.

GMMs can assume several different forms, depending on the type of covariance matrices. The two most widely used are full and diagonal covariance matrices. The number of GMM parameters that have to be estimated from sample or training data is given by $\frac{M}{2}(K^2 + 3K + 2)$ for full matrices and $M(2K+1)$ for diagonal matrices.

The parameters of the GMM are commonly estimated via the EM algorithm [8]. The algorithm iterations produce a sequence of models with monotonically nondecreasing (log) likelihood values. Though the EM algorithm converges to a maximum likelihood it has a few drawbacks: it is a greedy algorithm and may converge to a local maximum and not the global maximum. GMMs produced by the EM algorithm are, consequently, sensitive to initialization. We use the *k-means* algorithm [9] to initialize the GMM parameters.

2.2 MMSE Estimation Using GMMs

The goal in GMM-based MMSE estimation is to find a mapping or regression function, $\hat{f}(\mathbf{x})$, that minimizes the mean squared error, ε_{MSE}, between predictor variables ($\mathbf{x}$) and the target variable (y), where

$$\varepsilon_{MSE} = E[(y - \hat{f}(\mathbf{x}))^2]. \qquad (2)$$

It is known that the mean squared error (2) is minimized when $\hat{f}(\mathbf{x}) = E[y|\mathbf{x}]$, the conditional expectation of the target variable, given the predictor vector.

GMM-based estimators rely on modelling the joint density of the K-dimensional predictor variables with the target

variable using (1) with $\mathbf{u} = [y, \mathbf{x}]^T$. The covariance matrix of the i^{th} GMM component becomes

$$\mathbf{\Sigma}_i = \begin{pmatrix} \Sigma_i^{yy} & \mathbf{\Sigma}_i^{yx} \\ \mathbf{\Sigma}_i^{xy} & \mathbf{\Sigma}_i^{xx} \end{pmatrix}.$$

Given the GMM parameters, the MMSE regression function is given by [1]

$$\hat{f}(\mathbf{x}) = E[y|\mathbf{x}] = \sum_{i=1}^{M} h_i(\mathbf{x})[\mu_i^y + \mathbf{\Sigma}_i^{yx}(\mathbf{\Sigma}_i^{xx})^{-1}(\mathbf{x} - \mu_j^x)]. \quad (3)$$

The above GMM estimator or GMM regressor function is a weighted sum of linear models, where the weight $h_i(\mathbf{x})$ is the probability that the i^{th} Gaussian component generated the vector $\mathbf{x}$ and given by

$$h_i(\mathbf{x}) = \frac{\dfrac{\alpha_i}{|\mathbf{\Sigma}_i^{xx}|^{1/2}} e^{\left(-\frac{1}{2}(\mathbf{x}-\mu_i^x)^T(\mathbf{\Sigma}_i^{xx})^{-1}(\mathbf{x}-\mu_i^x)\right)}}{\sum_{k=1}^{M} \dfrac{\alpha_k}{|\mathbf{\Sigma}_k^{xx}|^{1/2}} e^{\left(-\frac{1}{2}(\mathbf{x}-\mu_k^x)^T(\mathbf{\Sigma}_k^{xx})^{-1}(\mathbf{x}-\mu_k^x)\right)}}. \quad (4)$$

If the covariance matrices are restricted to be diagonal, (3) simplifies to

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{M} h_i(\mathbf{x})\mu_i^y. \quad (5)$$

3. FEATURE SELECTION

The primary objective in the feature selection problem is to find, amongst n candidate feature variables, a subset of variables $\{x_1, \ldots, x_m\}, m \leq n$, and a mapping $\hat{f}(x_1, \ldots, x_m)$, such that $\hat{f}$ yields a good estimate of the response variable y. Presently, CART [5] and MARS [6] have been used as feature selection algorithms and/or regressors in many diverse problems, including speech quality prediction [10]. They are used here to benchmark our proposed feature selection algorithm, which is described below.

3.1 Proposed Method

It is argued in [1] that the GMM estimator has interesting relations to models such as CART and MARS in the sense that the mixture of Gaussians competitively partitions the feature space and learns a linear regression surface on each partition. Thus, it seems evident that one should use the GMM estimator to sift out the most relevant variables. Here we propose a sequential feature selection algorithm that progressively constructs $\hat{f}$ using (3) or (5).

The proposed algorithm starts with an empty feature set and features from a candidate feature set are added to the set progressively. To determine which candidate feature to add, the algorithm tentatively adds to the current feature set one feature that is not already selected to form an augmented feature set. The joint density of the target variable and the augmented feature set is modelled with a GMM, with model parameters $\lambda = (\alpha, \mu, \mathbf{\Sigma})$ estimated using the EM algorithm. The accuracy of the GMM estimator using λ is then calculated. The above is repeated for every candidate feature and corresponding GMM. The candidate feature that produces the least regression error is admitted into the current feature set to form an updated feature set. The algorithm stops when the desired number of features has been selected.

It is worth mentioning that for each candidate feature the best number of Gaussian components, M, in (1) can be determined by checking different values of M. With a corresponding increase in computational complexity, multiple features can also be tested and selected per iteration. Using the notation "EM" to stand for GMM parameter estimation via the EM algorithm, $\hat{f}_k$ for the regression function with k variables, and D for the desired number of features, the algorithm can be summarized as follows:

Initialization: Let $I = \{1, \ldots, n\}$, $S = \emptyset$, $k = 1$;

Step 1: $\lambda_i \leftarrow \text{EM}(y, S \cup \{x_i\}), \forall i \in I$;

Step 2: $i_k = \arg\min_{i \in I} \sum_j (y_j - \hat{f}_k(S \cup \{x_i\}|\lambda_i))^2$;

Step 3: $I \leftarrow I - \{i_k\}$, $S \leftarrow S \cup \{x_{i_k}\}$, $k \leftarrow k + 1$.

Go to step 1 if $k < D$, else stop.

Also note that for full covariance matrices the number of parameters that need to be estimated scales quadratically with the feature space dimension. When dealing with limited data, as in our case, severe problems arise due to singularities and local maxima in the log-likelihood function. Here we avert ill-conditioning by adding a small diagonal matrix, namely $\varepsilon I_{n \times n}$, to each covariance matrix in each M-step iteration of the EM algorithm. Typically, the optimal value for ε is not known *a priori*. A simple procedure used here is to vary ε over a range of values and choose the value that leads to the best performance on the validation data set. We varied ε from 10^{-2} to 10^{-9} and the value that led to best performance was $\varepsilon = 10^{-9}$.

We dedicate the next section to testing the proposed feature selection algorithm on a GMM-based speech quality estimation task. We compare with GMM estimators trained on features selected by CART or MARS. We present results for both diagonal and full covariance GMMs.

4. EXPERIMENT SETUP

The GMM for speech quality estimation is built on perceptual feature variables obtained by classifying perceptual distortions under a variety of contexts, as proposed in [10]. A total of 206 candidate features are extracted per speech file pair. In [4], the top-5 most important feature variables as ranked by CART and MARS are used for training GMM estimators. The target variable is the subjective quality rating represented by the Mean Opinion Score (MOS) [11], which falls between 1 and 5.

We compare GMM estimators trained on features selected by our proposed feature selection algorithm to estimators trained on features selected by CART or MARS. Thirteen MOS labelled speech databases are used, containing a total of 5864 speech files. We use 10-fold cross validation to provide some robustness in the performance evaluation. Estimation performance is assessed by the correlation (R) between subjective and estimated MOS. MOS measurement accuracy is assessed using the root-mean-square error ($RMSE$).

For this experiment we check all permissible values of M at each iteration. To allow comparisons with [4] we choose

the top-5 features and restrict $M \leq 5$ in order to maintain an adequate training ratio (ratio between the number of parameters that have to be estimated during the training phase and the total number of files in the training set) of 37 for full covariance matrices and 81 for diagonal matrices.

Let M_i be the number of Gaussian components chosen in iteration i of the proposed algorithm, it was found that the following combinations were often selected throughout the ten cross validation trials:

- Diagonal: $M_1 = 4$, $M_2 = M_3 = M_4 = M_5 = 5$;
- Full: $M_1 = 2$, $M_2 = 3$, $M_3 = M_4 = 4$, $M_5 = 5$.

Note that for the five algorithm iterations (D=5) used in this experiment the number of Gaussian components either increases or stays the same as the algorithm progresses. As expected, full covariance GMMs use fewer Gaussian components at the beginning, and the number of components increases with the number of features.

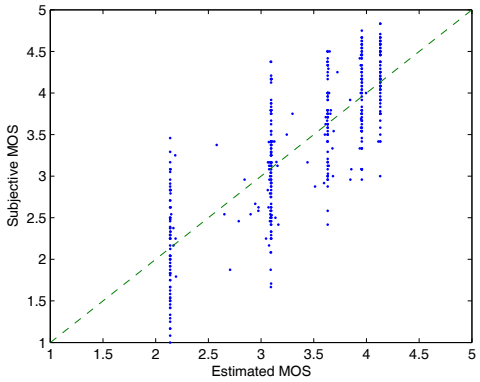
For diagonal GMMs, the most important selected features amount to distortions of severely distorted voiced and unvoiced frames. For full covariance GMMs, weighted mean distortion and root-mean distortion of voiced and unvoiced frames are often selected as being the most important (refer to [4] for a more detailed description of the feature variables). Tables 1 and 2 compare performance figures for a five-component GMM estimator designed using the proposed algorithm to that of an estimator designed using CART or MARS, for diagonal and full covariance matrices, respectively. The column "$\%\uparrow$" indicates percentage improvement in R found by using features selected with our proposed method. The percentage improvement is given by

$$\%\uparrow R = \frac{R_{new} - R_{old}}{1 - R_{old}} \times 100\%$$

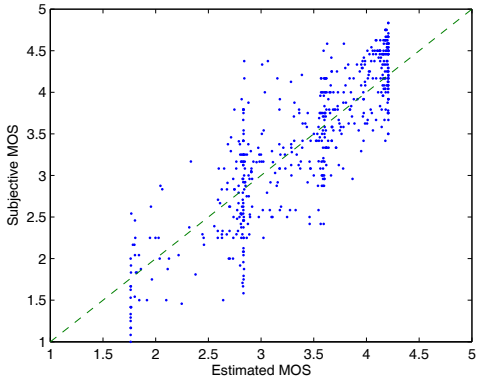
where R_{new} and R_{old} are the correlation obtained using the proposed method and using CART or MARS, respectively. The improvement indicates percentage reduction of the gap to perfect correlation. In turn, column "$\%\downarrow$" indicates percentage reduction in $RMSE$.

As can be seen, our proposed algorithm outperforms both benchmark algorithms. For diagonal GMM estimators we see an average improvement in R of 26.95% and 38.94 %, and an average decrease in $RMSE$ of 13.93% and 24.16% when compared to CART and MARS, respectively. An average improvement in R of 31.10% and 20.01%, and an average decrease in $RMSE$ of 19.07% and 11.96% is achieved for full GMM estimators. Also note that an average improvement over PESQ [12], the current "state-of-art" speech quality prediction algorithm, of 26.11% in R and 18.04% in $RMSE$ is attained with our full GMM estimators.

One of the drawbacks of using CART and MARS for speech quality estimation was highlighted in [4] and consisted in the fact that features selected by the data mining algorithms had significant correlation amongst them. Diagonal covariance GMM estimators, consequently, presented only modest performance figures and this was attributed to the fact that the use of a small number of diagonal Gaussian components ($M = 5$) was insufficient to model or compensate for the correlation between features. By looking at Tables 1 and 2 we see that CART selected features outperform MARS selected features with diagonal GMM estimators. This does not hold true when using full GMM estimators, suggesting that MARS selected features are more correlated.



(a)



(b)

Figure 1: Subjective MOS *versus* Estimated MOS for (a) CART and (b) diagonal GMM selected features.

Furthermore, in [4] a prominent vertical alignment of points in the subjective MOS *versus* estimated MOS scattered plots (vide Fig.1(a)) suggested that full covariance GMM estimators were needed in order to predict the residual variation in subjective MOS. If one insists on using diagonal GMM, the problem is mitigated by using the proposed feature selection algorithm, as shown in Fig.1(b). Note that the vertical alignment of points is considerably less accentuated than CART, reflecting the performance improvements shown in Table 1. For MARS, the scattered plot is similar to CART and is omitted for brevity.

Observe in Figure 1(a) the five discrete estimated MOS values associated with the five diagonal Gaussian components (see (5)) are prominently indicated by the horizontal locations of the vertical clusters. In this case, the weights in (5) serve the sole purpose of switching between the five discrete values.

5. CONCLUSION

A feature selection algorithm is proposed for estimation based on Gaussian mixture models. The algorithm is targeted to applications that make use of GMM estimators, as features are selected to minimize squared GMM estimation errors. An experiment consisting of predicting the subjective quality rating of speech is performed. Simulation results show that GMM estimators designed using our proposed algorithm outperform two benchmark algorithms. Furthermore, we have also shown that features selected by the proposed algorithm

Table 1: Performance Comparison: diagonal covariance matrices

Cross Validation	Proposed		CART				MARS			
Trials	*R*	*RMSE*	*R*	%↑	*RMSE*	%↓	*R*	%↑	*RMSE*	%↓
Trial 1	0.8578	0.4390	0.8083	25.82	0.5016	14.25	0.7926	31.44	0.5206	18.58
Trial 2	0.8539	0.4623	0.8216	18.11	0.5036	8.93	0.7465	42.37	0.5577	20.63
Trial 3	0.8530	0.4479	0.7972	27.51	0.5068	13.15	0.7903	29.90	0.5140	14.75
Trial 4	0.8732	0.4448	0.8206	29.32	0.4930	10.83	0.7661	45.79	0.5821	30.86
Trial 5	0.8416	0.4585	0.7937	23.22	0.5126	11.79	0.6863	49.51	0.6151	34.15
Trial 6	0.8694	0.4266	0.8184	28.08	0.4903	14.93	0.7479	48.20	0.5709	33.82
Trial 7	0.8740	0.4305	0.8171	31.11	0.5111	18.72	0.8089	34.07	0.5243	21.78
Trial 8	0.8656	0.4409	0.8171	26.52	0.4996	13.31	0.8043	31.32	0.5159	17.01
Trial 9	0.8521	0.4623	0.7879	30.27	0.5400	16.80	0.8000	26.05	0.5255	13.67
Trial 10	0.8677	0.4341	0.8122	29.55	0.5061	16.58	0.7313	50.76	0.5919	36.35
Average				26.95		13.93		38.94		24.16

Table 2: Performance Comparison: full covariance matrices

Cross Validation	Proposed		CART				MARS			
Trials	*R*	*RMSE*	*R*	%↑	*RMSE*	%↓	*R*	%↑	*RMSE*	%↓
Trial 1	0.8931	0.3830	0.8404	33.02	0.4627	20.81	0.8694	18.15	0.4209	9.89
Trial 2	0.8917	0.4005	0.8498	27.90	0.4656	16.25	0.8816	8.53	0.4168	4.06
Trial 3	0.8835	0.3930	0.8452	24.74	0.4480	13.99	0.8651	13.64	0.4218	7.32
Trial 4	0.9023	0.3907	0.8648	27.74	0.4557	16.63	0.8923	9.29	0.4094	4.78
Trial 5	0.8852	0.3923	0.8322	31.59	0.4671	19.06	0.8336	31.01	0.4657	18.71
Trial 6	0.8919	0.3889	0.8498	28.03	0.4531	16.50	0.8742	14.07	0.4173	7.30
Trial 7	0.8953	0.3955	0.8538	28.39	0.4626	16.96	0.8900	4.82	0.4060	2.65
Trial 8	0.9075	0.3644	0.8574	35.13	0.4467	22.58	0.8749	26.06	0.4196	15.14
Trial 9	0.8963	0.3889	0.8329	37.94	0.4846	24.61	0.8553	28.33	0.4541	16.76
Trial 10	0.9047	0.3708	0.8498	36.55	0.4572	23.30	0.8229	46.19	0.4931	32.98
Average				31.10		19.07		20.01		11.96

are suitable for diagonal GMM estimators, which incur lower computational complexity.

REFERENCES

[1] Z. Ghahramani and M. I. Jordan, "Supervised learning from incomplete data via an EM approach," in *Adv. in Neural Inf. Proc. Systems*, vol. 6. Morgan Kaufmann Publishers, Inc., 1994, pp. 120–127.

[2] A. Kain and Y. Stylianou, "Stochastic modeling of spectral adjustment for high quality pitch modification," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Proc.*, vol. 2, June 2000, pp. 949–952.

[3] R. Martin, C. Hoelper, and I. Wittke, "Estimation of missing LSF parameters using Gaussian mixture models," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Proc.*, vol. 2, May 2001, pp. 729–732.

[4] T. H. Falk and W.-Y. Chan, "Feature mining for GMM-based speech quality measurement," in *Proc. of the 38th Asilomar Conf. on Signals, Systems and Computers*, Nov. 2004, pp. 2290–2294.

[5] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks, 1984.

[6] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–141, March 1991.

[7] M. Law, M. Figueiredo, and A. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1154–1166, Sept. 2004.

[8] A. Dempster, N. Lair, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[9] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

[10] W. Zha and W.-Y. Chan, "Objective speech quality measurement using statistical data mining," *EURASIP Journal of Applied Signal Proc.*, to appear in 2005.

[11] ITU-T Rec. P.830, "Subjective performance assessment of telephone-band and wideband digital codecs," 1996.

[12] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.